# Web Mining for Multimedia Question Answering

Minerva Project Briefing

Yiming Yang

School of Computer Science

Carnegie Mellon University

# The Team

- Yiming Yang (PI)
- Jaime Carbonell (Co-PI)
- Ulas Bardak (grad student)
- Ashwin Tengli (grad student)
- Bryan Kisiel (programmer)
- Denise Noyes (undergrad)

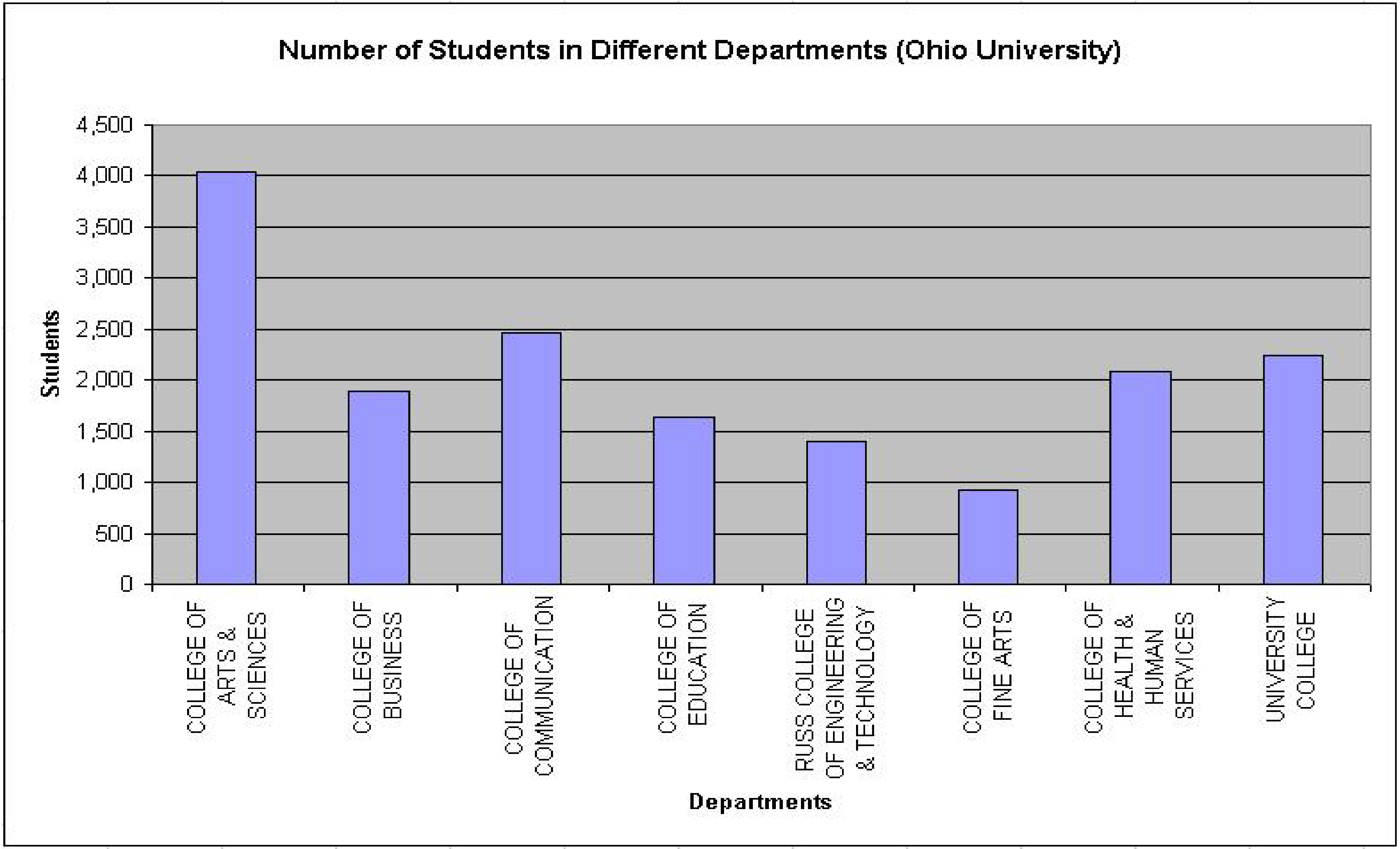
# Primary Aims

- **Web Mining**
  - Statistics in tables, graphs, structured records
  - About Named Entities (criminals, universities, cities, …)
  - Patterns of statistics over time
- **Question Answering**
  - Structured databases (of extracted information)
  - Semi-structured questions
  - Similarity-based ranking of tables, graphs, records

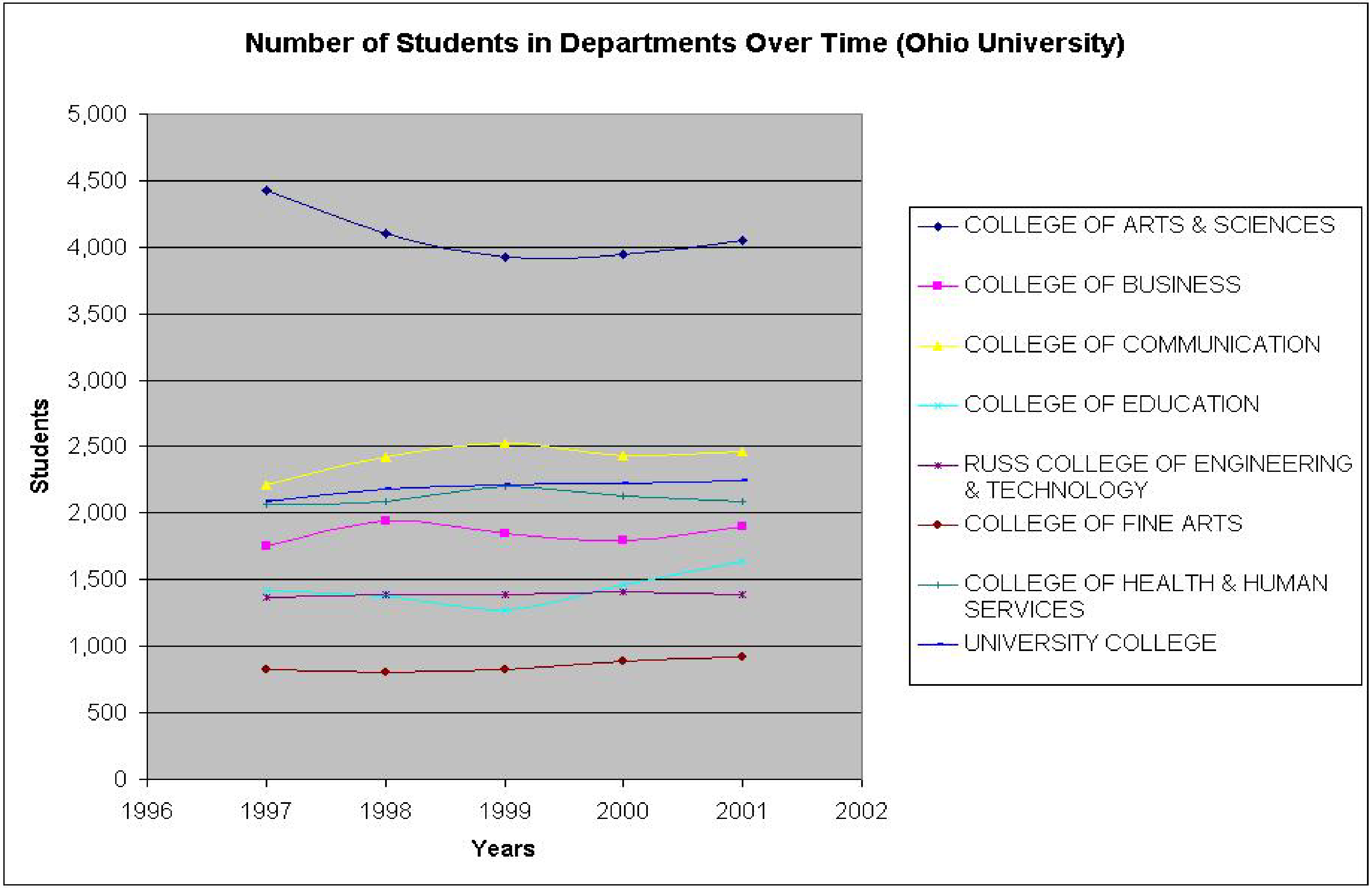
# Sample Questions

- **Student Demographics**
  - What are the main changes, if any, in the past decade?
  - Have other universities exhibited a similar trend?
  - What is the distribution of students in university X by departments?

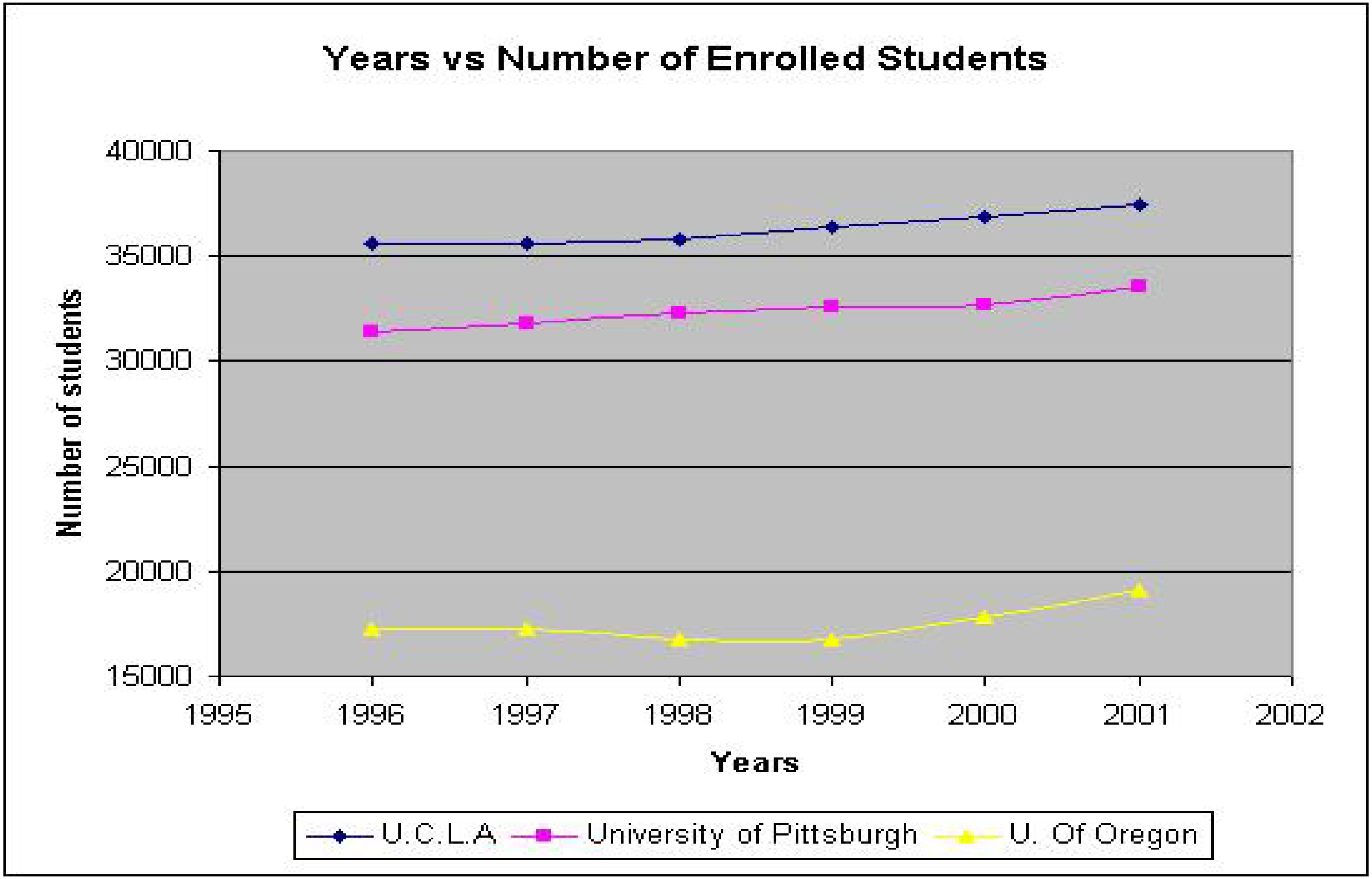
# A picture is worth a thousand of words



# A picture is worth a thousand of words



# A picture is worth a thousand of words



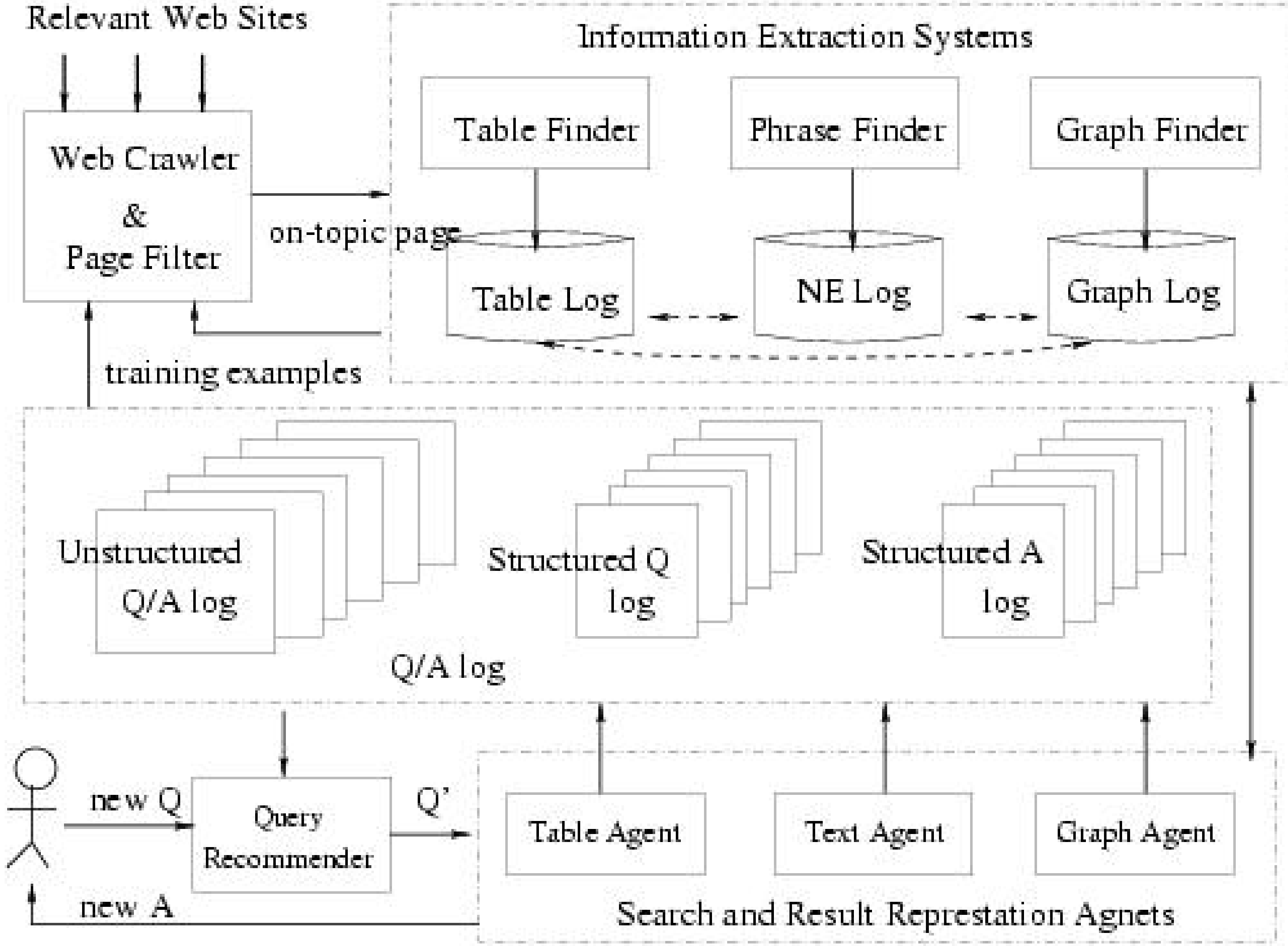
# Rationale

- Statistic-based QA is beneficial for certain types of questions
- Web pages contains rich statistical information
- Information extraction techniques need to be developed and improved Web mining
- Comparative image analysis (on curves, graphs) should be investigated for QA based on statistics

# What is novel?

- Mining the Web for distributed statistical information
- Answering questions using statistics in a tabular or graphical form
- Developing meaningful similarity metrics for comparing curves or temporal trends

**Figure 1. Overview of the Multimedia Q/A System**



# Next …

- Q/A
  - Question Templates
  - Answer Formulation
    - ***Comparing Curves***
  - Query Relaxation
- Web Mining
  - Focused crawling of ***invisible*** Web sites
    - Using Named Entities with statistical weights
  - Wrapper induction for ***different*** web sites
    - Supervised learning
  - Information extraction for ***tabular*** data

# Question Templates

- Three domains chosen -- universities, criminals, properties
- Templates defined for each domain
- Assumed that different domains will have different templates
- Allowing questions to relate more than one domain at the same time
- Chose XML to formulate questions

# Representation - Tables

**CARNEGIE MELLON UNIVERSITY**

B. ENROLLMENT AND PERSISTENCE

**B1.** Institutional Enrollment--Men and Women Provide numbers of students reported on IPEDS Fall Enrollment Survey 1999 as of the institution's official fall reporting date or as of October 15, 1999. Refer to IPEDS EF-1 Part A or IPEDS EF-2 Part A (undergraduates only) survey.

| | FULL-TIME | | | PART-TIME | | |
|---|---|---|---|---|---|---|
| | Men (IPEDS col. 15) | Women (IPEDS col. 16) | IPEDS line | Men (IPEDS col. 15) | Women (IPEDS col. 16) | IPEDS line |
| **Undergraduates** | | | | | | |
| Degree-seeking, first-time freshmen | 768 | 486 | line 1 | 0 | 0 | line 15 |
| Other first-year, degree-seeking | 27 | 16 | line 2 | 1 | 1 | line 16 |
| All other degree-seeking | 2446 | 1304 | lines 3-6 | 60 | 27 | lines 17-20 |
| *Total degree-seeking* | 3241 | 1806 | | 61 | 28 | |
| All other undergraduates enrolled in credit courses | 1 | 3 | line 7 | 63 | 59 | line 21 |
| *Total undergraduates* | **3242** | **1809** | line 8 | **124** | **87** | line 22 |
| **First-professional** | | | | | | |
| First-time, first-professional students | 0 | 0 | line 9 | 0 | 0 | line 23 |
| All other first-professionals | 0 | 0 | line 10 | 0 | 0 | line 24 |
| *Total first-professional* | 0 | 0 | | 0 | 0 | |
| **Graduate** | | | | | | |
| Degree-seeking, first-time | 703 | 300 | line 11 | 167 | 77 | line 25 |
| All other degree-seeking | 1021 | 368 | line 12 | 339 | 199 | line 26 |
| All other graduates enrolled in credit courses | 0 | 0 | line 13 | 0 | 0 | line 27 |
| Total graduate | **1724** | **668** | | **506** | **276** | |

Total all undergraduates (IPEDS sum of lines 8 and 22, cols. 15 and16): **5136***
*** Total does not include non-degree seeking students.**

Total all graduate and professional students (IPEDS sum of lines 14and 28, cols. 15 and 16): **3174**

## Templates - CDS

- Name of school
- Year covered
- Number of fulltime freshman males accepted
- Number of fulltime freshman females accepted
- Number of parttime freshman males accepted
- Number of parttime freshman females accepted
- Total number of students
- Number of nonresident aliens accepted
- Number of black,non-hispanics accepted
- Number of American Indian or Alaskans accepted
- Number of Asian or Pacific Islanders accepted
- Number of Hispanics accepted
- Number of White, non-hispanics accepted
- Number of Students who submitted SAT scores
- Percent of freshman with SAT1 verbal 700-800
- Percent of freshman with SAT1 verbal 600-699
- Percent of freshman with SAT1 verbal 500-599
- Percent of freshman with SAT1 verbal 400-499
- Percent of freshman with SAT1 verbal 300-399
- Percent of freshman with SAT1 verbal 200-299
- Percent of freshman with SAT1 math 700-800
- Percent of freshman with SAT1 math 600-699
- Percent of freshman with SAT1 math 500-599
- Percent of freshman with SAT1 math 400-499
- Percent of freshman with SAT1 math 300-399
- Percent of freshman with SAT1 math 200-299
- School total cost per year

## Templates – General U. Info

- Name of College/University
- Mailing Address
- City/State/Zip
- Main Phone
- Homepage
- Source of control (public/private/proprietary)
- Classification (coed,men,women)
- Degrees Offered

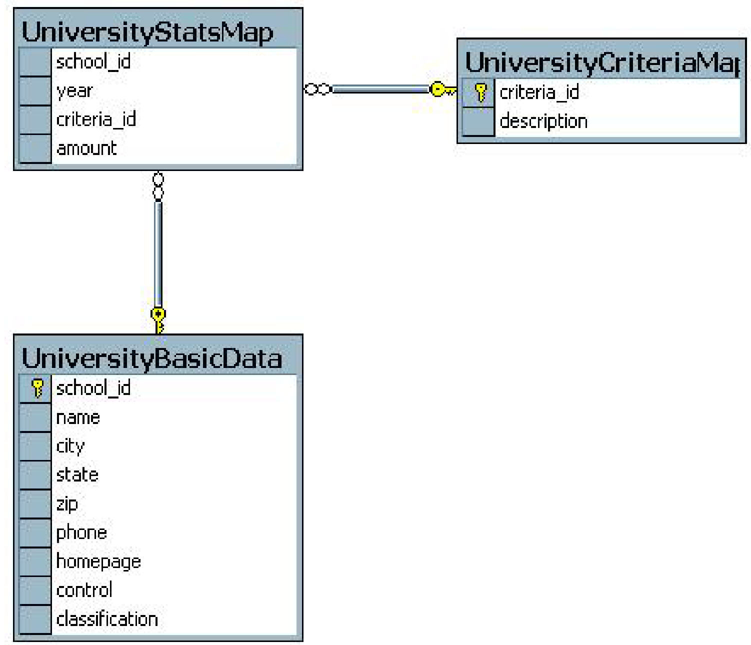
# Representation - Tables

## Templates - CDS

- Name of school
- Year covered
- Number of fulltime freshman males accepted
- Number of fulltime freshman females accepted
- Number of parttime freshman males accepted
- Number of parttime freshman females accepted
- Total number of students
- Number of nonresident aliens accepted
- Number of black,non-hispanics accepted
- Number of American Indian or Alaskans accepted
- Number of Asian or Pacific Islanders accepted
- Number of Hispanics accepted
- Number of White, non-hispanics accepted
- Number of Students who submitted SAT scores
- Percent of freshman with SAT1 verbal 700-800
- Percent of freshman with SAT1 verbal 600-699
- Percent of freshman with SAT1 verbal 500-599
- Percent of freshman with SAT1 verbal 400-499
- Percent of freshman with SAT1 verbal 300-399
- Percent of freshman with SAT1 verbal 200-299
- Percent of freshman with SAT1 math 700-800
- Percent of freshman with SAT1 math 600-699
- Percent of freshman with SAT1 math 500-599
- Percent of freshman with SAT1 math 400-499
- Percent of freshman with SAT1 math 300-399
- Percent of freshman with SAT1 math 200-299
- School total cost per year

## Templates – General U. Info

- Name of College/University
- Mailing Address
- City/State/Zip
- Main Phone
- Homepage
- Source of control (public/private/proprietary)
- Classification (coed,men,women)
- Degrees Offered

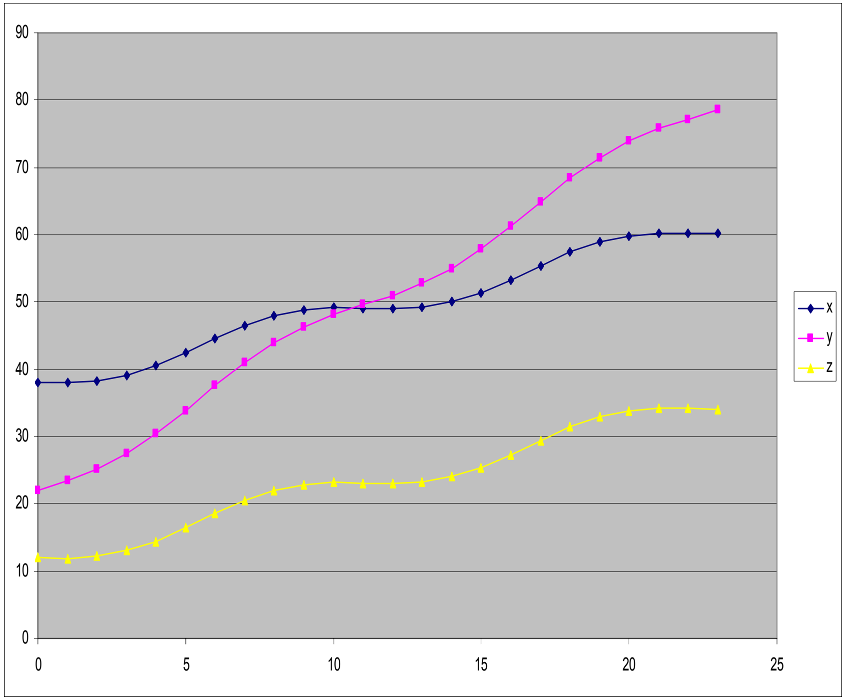


# Comparing Curves

- Once we retrieve the data for a query, we can fit a curve and compute its trend.
- We need to figure out which trends are most "similar"
  - A lot of different approaches possible
  - Need testing to figure out if any is good enough
  - Or, if we need to define a new metric.

# Comparing Curves

- Let's start simple:



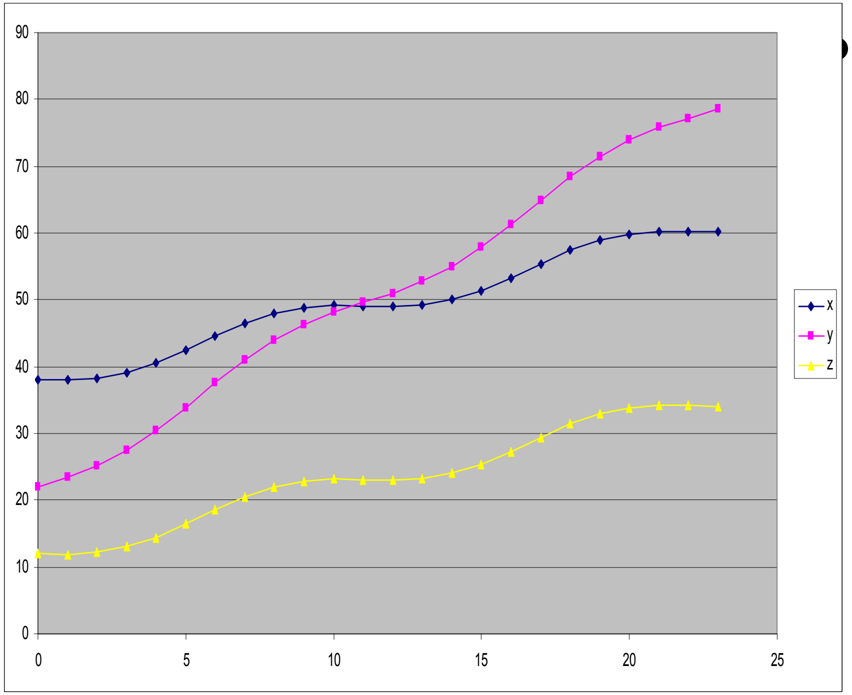

- All series have same number of data points.
  - 1-to-1 correspondence

$$\vec{x} = (x_1, x_2, \ldots, x_{24}), \vec{y} = (y_1, y_2, \ldots, y_{24}), \vec{z} = (z_1, z_2, \ldots, z_{24})$$

ARDA

# Comparing Curves

- Let's start simple:




- Approach 1:
  - Define dissimilarity by point wise comparison:

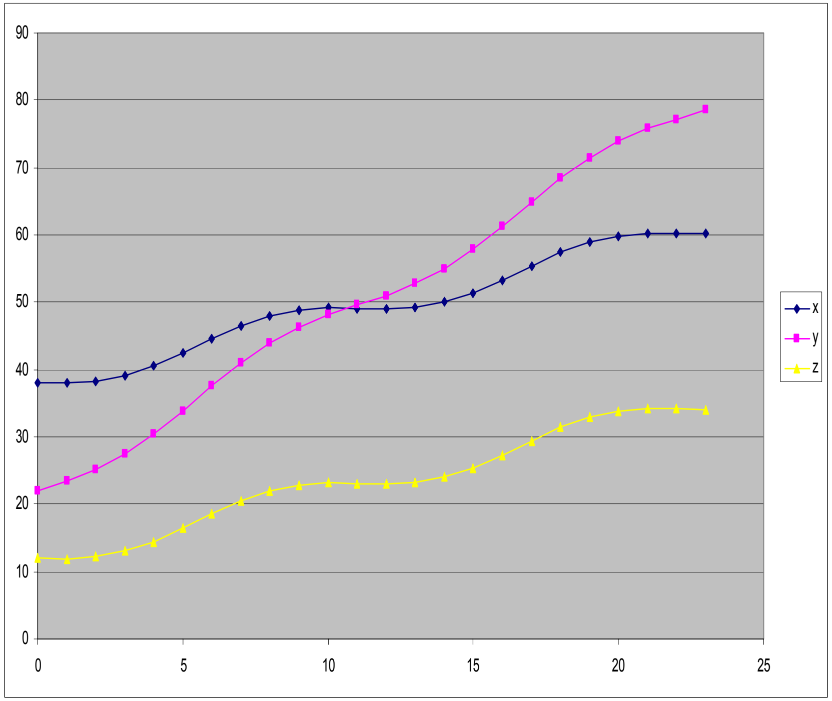
$$\mathrm{Dis1}(\vec{x}, \vec{y}) = \sum_{i=1}^{24} (x_i - y_i)$$

  - Picks curves *x* and *y*.

Clearly not what we want!

# Comparing Curves

- Let's start simple:




- Approach 2:
  - Define dissimilarity by difference in changes over consecutive points:

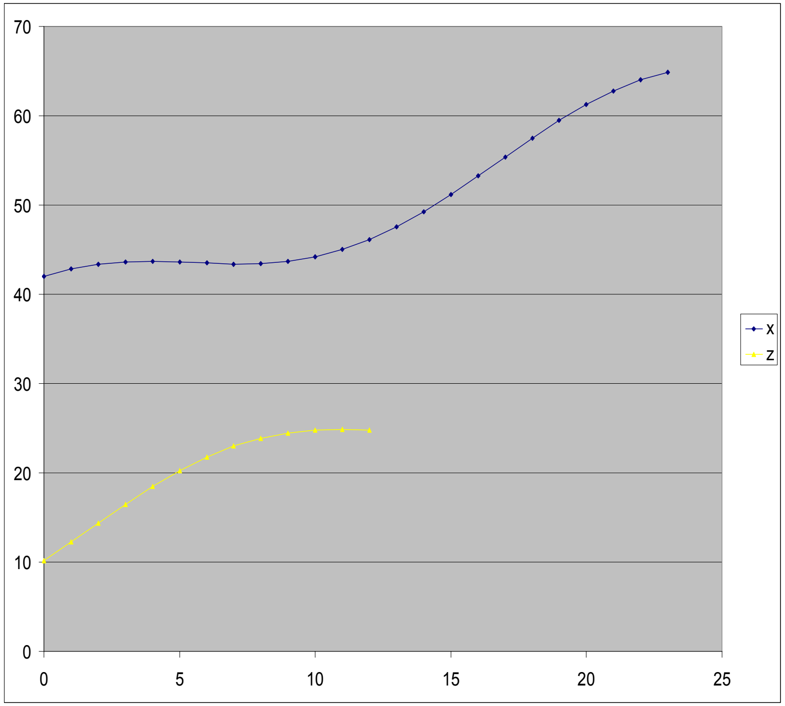
$$\mathrm{Dis2}(\vec{x}, \vec{y}) = \sum_{i=1}^{24} (\Delta x_i - \Delta y_i)$$

where $\Delta x_i = x_i - x_{i-1}$ and $\Delta y_i = y_i - y_{i-1}$

  - Picks curves *x* and *z*.

# Comparing Curves

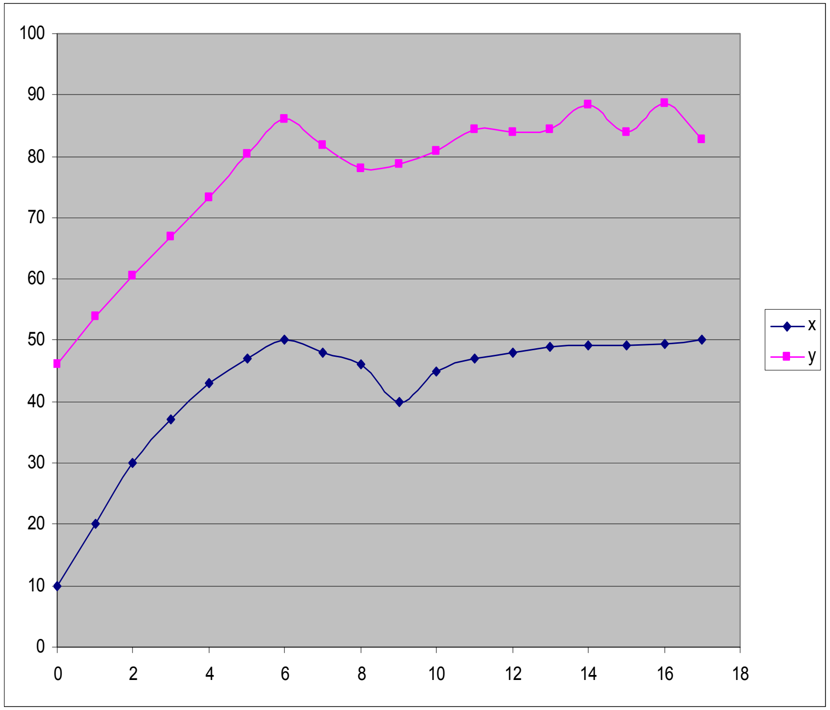
- Relaxing assumption – handling missing data:




- Need for filling in missing data inside the series – *interpolation*
- Need for filling missing data at the end of the series – *extrapolation*

# Comparing Curves

- A (relatively) complex approach:

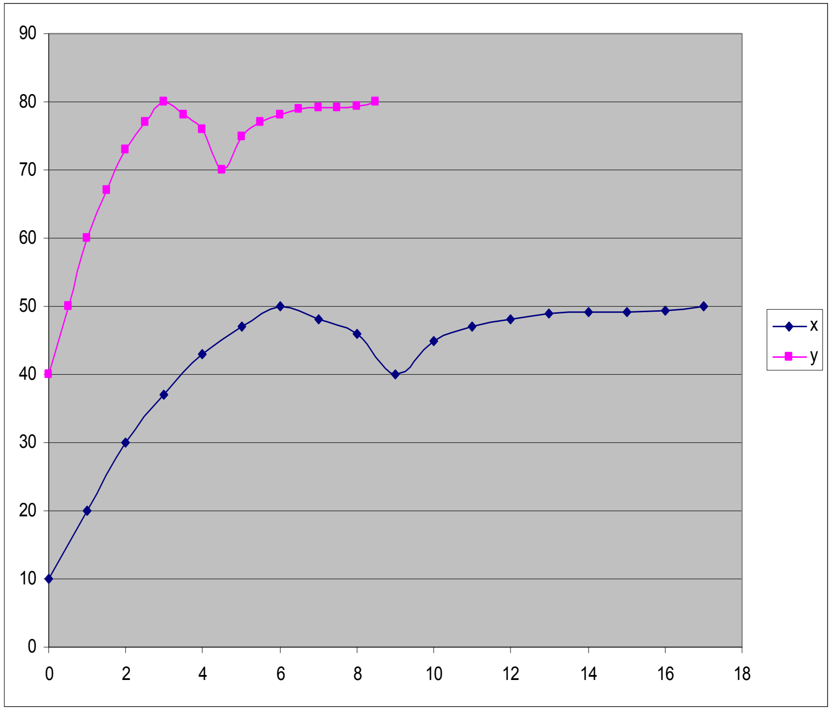



We can approximate each slope as a series of lines and compare those.

$$\text{Total Dissimilarity}(\vec{x},\vec{z}) = \sum_{i=1}^{total\_number\_of\_lines} \text{Dis1}(line_{i\text{ of }\vec{x}}, line_{i\text{ of }\vec{z}})$$

# Comparing Curves

- Linear renormalization of curves:

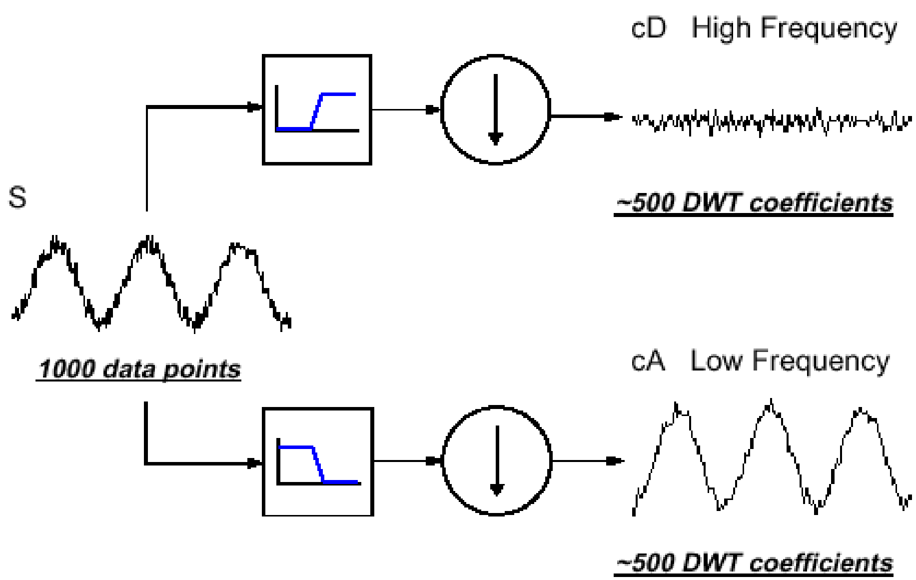



What if one curve only covers a part of the total time scale?

- Need to stretch so that we can compare with the target curve.
- Probably a good idea to punish the short curve though.

# Comparing Curves

- Taking analysis a step further:




We can use wavelets to decompose the 'signal'

- We get a "high frequency" and a "low frequency" component.
- For complicated graphs can reveal underlying relationships.
- Probably an overkill for basic graphs.

# Evaluation Plan

- **Data collections**
  - *University Students* corpus
  - *Property Ownership* corpus
  - *Criminal Records* corpus
- **What to Evaluate?**
  - End-to-end performance (black box)
    - Human annotated, paired test query-answer(s)
    - Precision, recall, F1, or MRR
  - Component performance (glass box)
    - Web mining module (quantity & quality of info)
    - Answer extraction module
      - Right graphs, tables for Q?
        - Similarity metrics for graphs & tables
    - User presentation/interaction interface

# Curve-Curve Similarity: Which is Best?

$$Sim_1(q,r \mid \vec{x},i) = \int q(\vec{x})dx_i - \int r(\vec{x})dx_i$$

$$Sim_2(q,r \mid \vec{x},i) = \frac{\partial q(\vec{x})}{\partial x_i} - \frac{\partial r(\vec{x})}{\partial x_i}$$

$$Sim_3(q,r \mid \vec{x},i) = \frac{\partial^2 q(\vec{x})}{\partial {x_i}^2} - \frac{\partial^2 r(\vec{x})}{\partial {x_i}^2}$$

# More on Curve Similarity

- Multiple methods
  - Analytic (wavelets or derivatives)
  - Inflection-point centric (Fink et al)
  - With displacement, scaling, embedding…
- Humans (analysts) are final arbiters
  - Maximal correlation of method with aggregate human judgments
  - Collect a set of graphs & similarity judgment corpus

# Concluding Remarks

- MINERVA is an Exploratory Project
  - Uncharted territories:
    - Graph or chart as query (and answer)
    - Mining for aggregate statistical data
  - Evaluation must track research
    - Component glassbox comes first
    - TREC-style in subsequent periods

# References

## Focused Crawling


1. Building Domain-Specific Search Engines with Machine learning Techniques., Andrew McCallum, Kamal Nigam, Jason Rennie and Kristie Seymore
2. Accelerated Focused Crawling through Online Relevance Feedback., Soumen Chakrabarti, Kunal Punera and Mallela Subramanyam., In WWW2002, Honolulu, Hawaii, USA.
3. Intelligent crawling on the World Wide Web with arbitrary predicates., C. C. Aggarwal, F. Al-Garawi, and P. S. Yu., In WWW2001, Hong Kong, May 2001. ACM.
4. Focused crawling: a new approach to topic-specic web resource discovery., S. Chakrabarti, M. van den Berg, and B. Dom., Computer Networks, 1999.
5. Using reinforcement learning to spider the web efficiently., J. Rennie and A. McCallum., In *ICML*, 1999.
6. Evaluating topic-driven Web crawlers., F. Menczer, G. Pant, M. Ruiz, and P. Srinivasan., In *SIGIR*, New Orleans, Sept. 2001.
7. Adaptive retrieval agents: Internalizing local context and scaling up to the Web., F. Menczer and R. K., In *Machine Learning*, 39(2/3):203-242, 2000.
8. Focused crawling using context graphs., M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori., In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*.
9. Searching for arbitrary information in the WWW: The fish search for Mosaic., P. M. E. De Bra and R. D. J. Post., In *Second World Wide Web Conference '94: Mosaic and the Web*, Chicago, Oct. 1994.
10. Efficient crawling through URL ordering., J. Cho, H. Garcia-Molina, and L. Page., In *7th World Wide Web Conference*, Brisbane, Australia, Apr. 1998.
11. The Web as a Graph: Measurement, Models and Methods., J. Kleinberg, R. Kumar, P. Raghavan, S. Rajgopalan and A.S. Tomkins., In Proc. 5th Annual Intl. Conf. Computing and Combinatorics, COCOON, 1999.

# References

## Table Detection and Extraction


12. A Machine Learning Based Approach for Table Detection on The Web., Yalin Wang and Jianying Hu., In WWW2002, Honolulu, Hawaii, USA.
13. Mining tables from large scale html texts., H.-H. Chen, S.-C. Tsai, and J.-H. Tsai., In *Proc. 18th International Conference on Computational Linguistics*, Saabrucken, Germany, July 2000.
14. Why table ground-truthing is hard., J. Hu, R. Kashi, D. Lopresti, G. Nagy, and G. Wilfong., In *Proc. 6th International Conference on Document Analysis and Recognition (ICDAR01)*, Seattle, WA, USA, September 2001.
15. Layout and language: Challenges for table understanding on the web., M. Hurst., In *Proc. 1st International Workshop on Web Document Analysis*, Seattle, WA, USA, September 2001.
16. A method to integrate tables of the world wide web., M. Yoshida, K. Torisawa, and J. Tsujii., In *Proc. 1st International Workshop on Web Document Analysis*, Seattle, WA, USA, September 2001.
17. A graph-based table recognition system., M. Rahgozar anf R. Cooperman., In Proc. Of Document Recognition III, SPIE, San Jose, California, 1996.

# References


## Mining the Hidden Web

18. Crawling the Hidden Web., Sriram Raghavan and Hector Garcia-Molina., Proceedings of the Twenty-seventh International Conference on Very Large Databases, 2001.
19. Automatic information discovery from the Invisible Web., King-Ip Lin and Hui Chen., Proceedings of the The International Conference on Information Technology: Coding and Computing (ITCC'02), 2002.
20. The Deep Web:Surfacing Hidden Value., http://www.completeplanet.com/Totorials/DeepWeb/

## Wrapper Induction

21. RoadRunner: Towards Automatic Data Extraction from Large Websites., Valter Crescenzi, Giansalvatore Mecca and Paolo Merialdo
22. Wrapper Induction: Efficiency and expressiveness, Nicholas Kushmerick
23. Generating finite-state transducers for semistructured data extraction from the web., C. Hsu and M. Dung., Information Systems, 1998.
24. A hierarchical approach to wrapper induction., I. Muslea, S. Minton and C. A. Knoblock., In Proc. Of Autonomous Agents, 1999.
25. NoDoSE- A tool for semi-automatically extracting structured and semi structured data from text documents., B. Adelberg., In SIGMOD’98
26. Extracting semi structured data through examples., B. A. Riberio-Neto., In CIKM’99
27. A conceptual-modeling approach to extracting data from the web., D.W. Embley, D. M. Campbell, Y. S. Jiang, S.W. Liddle, Y. Ng, D. Quass and R. D. Smith., In ER’98.

# References

## Curve analysis


28) Ian Kaplan, Applying the Haar Wavelet Transform to Time Series Information, http://www.bearcave.com/misl/misl_tech/wavelets/haar.html

29) W. W. Chu, H. Yang, K. Chiang, M. Minock, G. Chow, C. Larson, CoBase: A Scalable and Extensible Cooperative Information System, Journal of Intelligent Information Systems, 1996

30) D. Rafiei and A. Mendelzon. Similarity-based queries for time series data. SIGMOD Record (ACM Special Interest Group on Management of Data), 26(2):13--25, May 1997.